姓名和学号：李胜志 2210180232

章节名称：分类决策树的基本原理

知识目标：掌握决策树用于分类任务的基本原理；掌握构建决策树的流程。

能力目标：能够使用决策算法训练模型，构造决策树。

素质目标：关注国家资讯，增强民族意识，培养爱国主义精神；理解决策树算法的基本原理，培养勇为人先的精神。

知识重点：使用决策算法训练模型，构造决策树。

知识难点：选择最佳特征和分割点。

分类决策树的基本原理：

分类决策树是一种以树形数据结构来展示决策规则和分类结果的模型。它主要用于数据挖掘技术中的分类分析。

决策树由根节点、内部节点和叶节点组成。每个非叶节点代表一个特征属性的测试，每个分支代表这个特征属性在某个值域上的输出，而每个叶节点存放一个类别。从根节点到叶节点的路径对应了一个判定的测试序列。

构建分类决策树的流程：

1. 从根节点开始，将所有数据划分到一个节点。
2. 判断数据集是否需要继续划分。如果数据为空集或样本都属于同一类，则停止划分。
3. 如果需要继续划分，则选择最优的属性进行划分。如何选择最优属性是决策树的关键，通常使用信息增益作为标准。
4. 根据选定的属性生成新的节点，并继续对子节点进行判断和划分，直到所有节点都满足停止条件。
5. 生成最终的决策树。

例如：购买计算机的客户数据集（表 1），使用决策树算法训练模型，可构造如图 1所示的决策树。

表 1 购买计算机的客户数据集

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 客户编号 | 年龄 | 收入 | 是否为学生 | 信用情况 | 购买计算机情况 |
| 1 | 青年 | 高 | 否 | 一般 | 没有购买 |
| 2 | 青年 | 高 | 否 | 好 | 没有购买 |
| 3 | 中年 | 高 | 否 | 一般 | 已购买 |
| 4 | 老年 | 中 | 否 | 一般 | 已购买 |
| 5 | 老年 | 高 | 是 | 一般 | 已购买 |
| 6 | 老年 | 高 | 是 | 好 | 没有购买 |
| 7 | 中年 | 中 | 是 | 好 | 已购买 |
| 8 | 青年 | 中 | 否 | 一般 | 没有购买 |
| 9 | 青年 | 高 | 是 | 一般 | 已购买 |
| 10 | 老年 | 中 | 是 | 一般 | 已购买 |
| 11 | 青年 | 中 | 是 | 好 | 已购买 |
| 12 | 中年 | 中 | 否 | 好 | 已购买 |
| 13 | 中年 | 高 | 是 | 一般 | 已购买 |
| 14 | 老年 | 中 | 否 | 好 | 没有购买 |

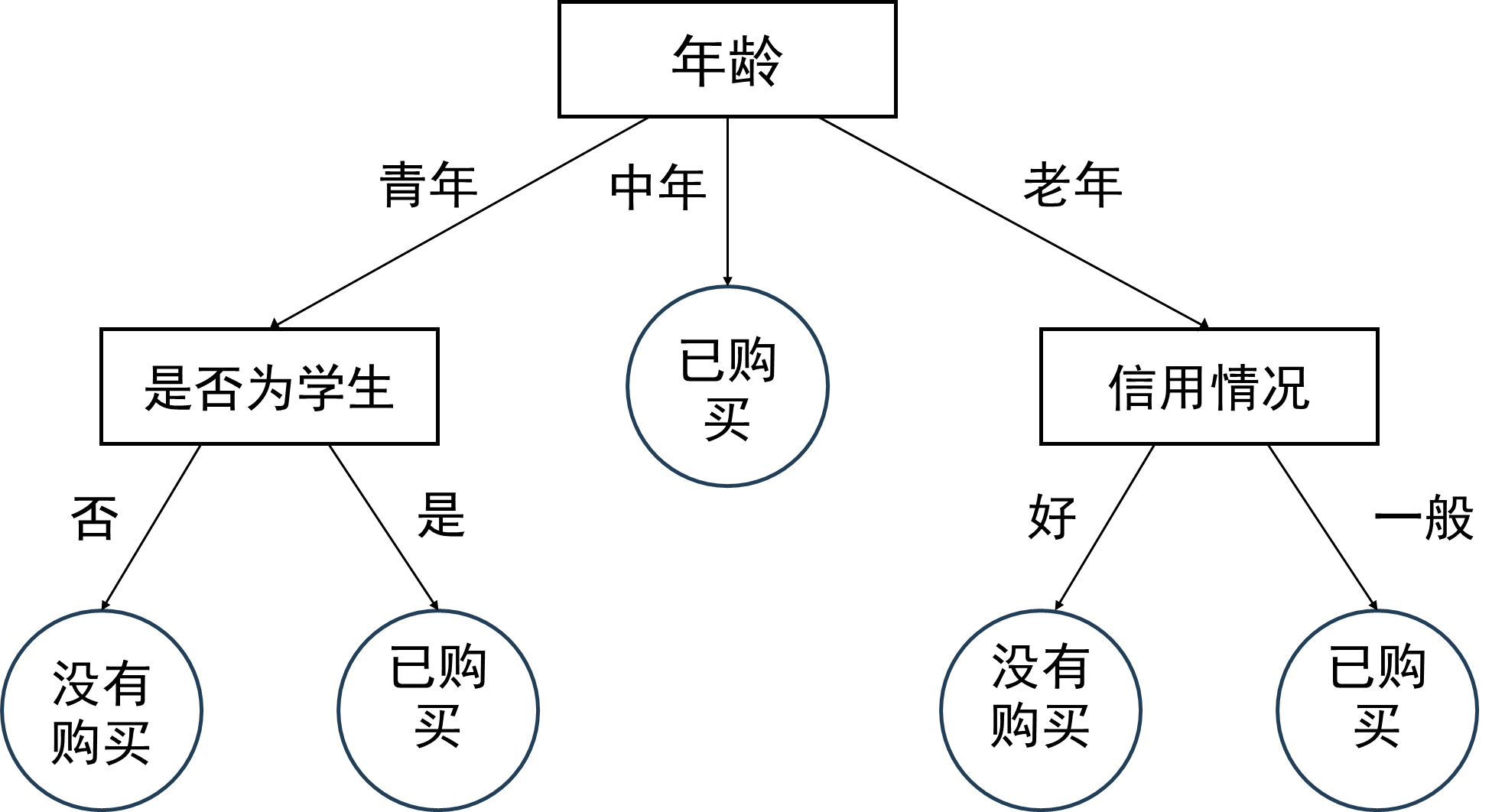


图 1分类决策树

从图6-1可以看出，决策树的根本节点和内部节点为数据集中的特征属性，叶节点为类别标签，根据特征属性的取值来判断进入哪一个分支。

决策树分类采用自顶向下的递归方式，在决策树内部节点进行属性值的比较，根据不同的属性值判断从该节点向下的分支，在叶节点上得到结论。所以，从决策树的根节点到叶节点的每一条路径都对应一条合取规则。例如，图6-1的决策树对应的决策规则为：

If年龄=青年^不是学生 Then 没有购买计算机

If年龄=青年^是学生 Then 已购买计算机

If年龄=中年 Then 已购买计算机

If年龄=老年^信用情况=好 Then 没有购买计算机

If年龄=老年^信用情况=一般 Then 已购买计算机

使用这些决策规则就可以对新的待测样本的类别进行判定。例如，新样本（老年，收入中等，不是学生，信用一般）的类别判定为已购买计算机。